CARLA GALVÃO FERNANDES

EFICIÊNCIA DA SELEÇÃO GENÔMICA COM BASE EM HAPLÓTIPOS AO LONGO DE GERAÇÕES, EM POPULAÇÕES COM NÍVEIS CONTRASTANTES DE DESEQUILÍBRIO DE LIGAÇÃO

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Genética e Melhoramento, para obtenção do título de *Magister Scientiae*.

Orientador: José Marcelo Soriano Viana

VIÇOSA - MINAS GERAIS
2021

Ficha catalográfica elaborada pela Biblioteca Central da Universidade Federal de Viçosa - Campus Viçosa

T

F363e
2021

Fernandes, Carla Galvão, 1994-
Eficiência da seleção genômica com base em haplótipos ao longo de gerações, em populações com níveis contrastantes de desequilíbrio de ligação / Carla Galvão Fernandes. - Viçosa, MG, 2021.
1 dissertação eletrônica (31 f.): il. (algumas color.).

Orientador: José Marcelo Soriano Viana.
Dissertação (mestrado) - Universidade Federal de Viçosa, Departamento de Biologia Geral, 2021.
Referências bibliográficas: f. 29-31.
DOI: https://doi.org/10.47328/ufvbbt.2021.181
Modo de acesso: World Wide Web.

1. Polimorfismo de nucleotídeo único. 2. Predição. 3. Confiabilidade. 4. Marcadores genéticos. I. Viana, José Marcelo Soriano, 1963-. II. Universidade Federal de Viçosa. Departamento de Biologia Geral. Programa de Pós-Graduação em Genética e Melhoramento. III. Título.

CDD 22. ed. 576.53

Bibliotecário(a) responsável: Alice Regina Pinto CRB6 2523

CARLA GALVÃO FERNANDES

EFICIÊNCIA DA SELEÇÃO GENÔMICA COM BASE EM HAPLÓTIPOS AO LONGO DE GERAÇÕES, EM POPULAÇÕES COM NÍVEIS CONTRASTANTES DE DESEQUILÍBRIO DE LIGAÇÃO

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Genética e Melhoramento, para obtenção do título de *Magister Scientiae*.

APROVADA: 28 de julho de 2021.

Assentimento:

Carla Galvão Fernandes
Autora

José Marcelo Soriano Viana
Orientador

Aos meus pais Carlos Antônio Fernandes e Rita Galvão, á Neide, minhas irmãs meus sobrinhos e ao meu namorado José Enrique.

DEDICO

AGRADECIMENTOS

Agradeço primeiramente a Deus e ao São Jorge que sempre estiveram comigo em minha determinação, seguindo e direcionando meus passos.

Á minha família. Aos meus pais Rita Galvão, Neném Braz e Neide, que sempre me incentivaram. Agradeço ao meu Padrinho Paulo, Madrinha Penha, Madrinha Mariane (irmã), Bebela (irmã) que sempre estiveram comigo me apoiando e dando grande exemplos como pessoas e profissionais.

À minha avó Alzira, vovó Leta (in memoriam), vovô Braz (in memoriam), Vovô Tãozinho (in memoriam) e Tia Dalena (in memoriam) pelo exemplo de vida que a cada dia me dão mais foças para lutar pelos meus novos sonhos.

Ao Hélcio que dedicou esses dois anos com grandes ensinamentos, me aturando com minhas reclamações e dificuldade, sendo um grande exemplo como homem e profissional.

Ao professor José Marcelo, que me recebeu de portas abertas no mestrado e com quem aprendi muito, foi ele que sempre acreditou em meu potencial e não me deixou desistir nos momentos de dificuldades.

Ao meu coorientador, Fabyano, que sempre confiou em mim e me guiou para essa conquista.

Ao meu coorientador Moyses, que confiou e cedeu o espaço em sua sala além de disponibilizar o servidor para que este momento se torne em realidade.

Ao professor Guilherme por ter aceitado o convite para participar da minha banca de defesa.

Ao Pedro do NuBioMol, que me instruiu com alguns comandos que foram essenciais para essa conquista.

O Dzianis Prakapenka, pela disponibilidade em tirar minhas dúvidas a respeito do software GVCHAP.

A Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), pela bolsa concedida e a UFV, pela oportunidade de fazer o mestrado. Sou grata ao Instituto de Biotecnologia Aplicada à Agropecuária (BIOAGRO) e à Diretoria de Tecnologia da Informação (DTI), Universidade Federal de Viçosa, por fornecer as instalações para a condução dos experimentos.

Aos amigos da sala 114 (BIOAGRO) e do Programa Milho-Pipoca, em que compartilhei grandes momentos, com muito aprendizado. Em especial, a Cynthia e ao

Matheus Ribeiro, que aguentaram minhas ladainhas e reclamações nos meus momentos de indecisões.

Aos meus amigos da UFV, Viçosa e parentes, que sempre me apoiaram e fizeram companhia. Muito obrigada Gustavo, Barbara, Pedro, “Manel’s Bar” e a todos que fazem dos meus dias mais felizes.

Agradeço também a toda família Espitia López, que me acolheram mesmo na distância com muito amor e grandes ensinamentos.

Gostaria de agradecer ao Felipe, por ter me orientado em minha primeira decisão de seguir os meus estudos após o ensino médio, que sem dúvida foi o primeiro grande passo para essa conquista.

À Universidade Federal de Viçosa, pela oportunidade de realizar a pós-graduação.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), pela concessão da bolsa de estudos.

Muito Obrigada!

RESUMO

FERNANDES, Carla Galvão, M.Sc., Universidade Federal de Viçosa, julho de 2021. **Eficiência da seleção genômica com base em haplótipos ao longo de gerações, em populações de diferentes níveis de desequilíbrio de ligação.** Orientador: José Marcelo Soriano Viana.

Estudos de seleção genômica comprovam que modelos baseados em blocos de haplótipos com população com alto desequilíbrio de ligação (LD) entre marcadores e locos de herança quantitativa (QTLs) possuem melhores acurácias de predição. Assim, faz-se necessário maiores investigações do comportamento das acurácias quando utilizamos populações com diferentes níveis de LD. O objetivo desse trabalho foi avaliar a eficiência da seleção genômica baseada em haplótipos, em um contexto de gerações avançadas de predição e alta densidade de marcas em duas populações com alto e baixo LD. Foram simulados fenótipos e genótipos de 5000 indivíduos não relacionados (fundadores) que originaram outros 5000 indivíduos nas próximas cinco gerações, sem seleção para as duas populações. Um total de 60000 polimorfismos de base única (SNPs) com densidade de um SNP a cada 0,01 cM em genoma com 10 cromossomos, foram utilizados dois softwares para comparar a eficiência nas acurácias, sendo eles o Software R e o GVCHAP. Na qual, em ambas adotamos haplótipos igual a 10 SNPs para essa comparação. Seguimos com o objetivo de comprar os acurácias adotando blocos de haplótipos com 4 e 10 SNPs em diferentes níveis de LD. A chamada de haplótipos foi feita para cada cromossomo após a obtenção da fase dos SNPs. A característica estudada nas predições, sob um modelo melhor preditor linear não-enviesado genômico (GBLUP), possuía herdabilidade em sentido amplo de 0,30. Para o processamento feito no software R para a população de baixo LD e com blocos de haplótipos igual a 10 SNPs. Obtivemos a acurácia variando de 0,30 a 0,25 para blocos de haplótipos e para analises com SNP individual variaram de 0,49 a 0,39. Reproduzindo as mesmas analises no software GVCHAP, as acurácias variaram de 0,49 a 0,13 para blocos de haplótipos com 10 SNPs e as predições baseadas em SNP individual foram de 0,57 a 0,20. Mantendo o uso do software GVCHAP, fizemos para um cenário de maior LD, com blocos de haplótipos com 4 SNPs, resultam em acurácias variando de 0,57 a 0,14, enquanto as predições baseadas em SNP individual variaram de 0,40 a 0,14 e as predições baseadas em blocos de haplótipos com 10 SNPs, as acurácias resultaram variando de 0,58 a 0,13. Os dois softwares demonstraram que, quando trabalhamos com população com menor LD o comportamento das acurácias são os mesmos, assim que para população de baixo LD além das acurácias reduzir ao longo de

gerações também contamos com melhores acurácias para análises de SNP individual comparado a blocos de haplótipos. No entanto, decidimos a seguir as analises no software GVCHAP pela facilidade em programação. As acurácias de ambas as populações decaíram ao longo das gerações independentemente do tamanho do bloco de haplótipos ou SNP individual. Porém, quando trabalhamos com população de maior LD, as análises por blocos de haplótipos são favorecidas, ou seja, obtivemos melhores acurácias quando comparados com as analises por SNP individual. Assim quando trabalhamos com a população de menor LD a acurácia de SNP individual superou as analises por haplótipos. No geral, as acurácias baseadas em haplótipos em população com maior LD supera as análises por marcas únicas. No entanto, em certo momento ao longo das gerações ocorre um decréscimo do LD, e as acurácias para SNPs tendem igualar ou superar as análises por haplótipos. Nosso estudo não dá suporte ao uso de haplótipos para a predição ao longo de muitas gerações ou para populações de baixo LD.

Palavras-chave: SNPs. Predição. Acurácia. Alta densidade de marcadores.

ABSTRACT

FERNANDES, Carla Galvão, M.Sc., Universidade Federal de Viçosa, July, 2021. **Efficiency of genomic selection based on haplotypes over generations, in populations with different levels of linkage disequilibrium**. Adviser: José Marcelo Soriano Viana.

Haplotype-based genomic selection studies with populations with high linkage disequilibrium (LD) between markers and quantitative trait loci (QTLs) show that haplotype-based models have better accuracy when compared to single-marker models. Thus, further research of the accuracy behavior is necessary when using groups with different levels of LD. The aim of this work was to evaluate the efficiency of genomic selection based on haplotypes in a context of advanced prediction generation and high marker density in two populations with high and low LD. Phenotypes and genotypes of 5000 unrelated (founders) were simulated, from which other 5000 individuals were derived in the next five generations, without selection for the two populations. A total of 60,000 single nucleotide polymorphisms (SNPs) with a density of one SNP every 0.01 cM was adopted in a genome with 10 chromosomes. Two software programs were used to compare the efficiency in accuracy, Software R and GVCHAP. In both, we adopted haplotypes equal to 10 SNPs for this comparison. We proceeded with the goal of buying the accuracies by adopting haplotype blocks with 4 and 10 SNPs at different LD levels. Haplotype calling was done for each chromosome after obtaining the phase of the SNPs. The trait studied in the predictions, under a best genomic unbiased linear predictor (GBLUP) model, had a broad sense heritability of 0.30. For the processing done in the R software for the low LD population and with haplotype blocks equal to 10 SNPs. We obtained accuracy ranging from 0.30 to 0.25 for blocks of haplotypes and for analyses with individual SNPs ranged from 0.49 to 0.39. Reproducing the same analyses in the GVCHAP software, the accuracies ranged from 0.49 to 0.13 for haplotype blocks with 10 SNPs and the predictions based on individual SNP ranged from 0.57 to 0.20. Maintaining the use of the GVCHAP software, we did for a higher LD scenario, with haplotype blocks with 4 SNPs, result in accuracies ranging from 0.57 to 0.14, while individual SNP-based predictions ranged from 0.40 to 0.14 and predictions based on haplotype blocks with 10 SNPs, the accuracies resulted ranging from 0.58 to 0.13. The two software's showed that when we work with populations with lower LD the behavior of the accuracies is the same, so that for low LD populations not only the accuracies reduce over generations but also, we have better accuracies for individual SNP analysis compared to haplotype blocks. However, we decided to follow the analysis in the GVCHAP software for the ease of programming. The accuracies of both populations

declined over generations regardless of the size of the haplotype block or individual SNP. However, when we worked with a population of higher LD, the analyses by haplotype blocks were favored, that is, we obtained better accuracy when compared to the analyses by individual SNP. Thus, when we worked with the lower LD population the individual SNP accuracy outperformed the haplotype analyses. Overall, haplotype-based accuracies in higher LD populations outperformed single-tag analyses. However, at some point along the generations there is a decrease in LD, and the accuracies for SNPs tend to equal or exceed the haplotype analyses. Our study does not support the use of haplotypes for prediction over many generations or for low LD populations.

Keywords: SNPs. Prediction. Accuracy. High marker density

LISTA DE TABELAS

LISTA DE FIGURAS

LISTA DE SIGLAS E ABREVIATURAS

bp	Pares de base
cM	Centímorgan
GBLUP	Modelo melhor preditor linear não-enviesado genômico
LD	Desequilíbrio de ligação
MAF	Frequência do alelo menor
QTL	Locos de herança quantitativa
SNP	Polimorfismo de base única

SUMÁRIO

1. INTRODUÇÃO

A predição genômica é uma maneira eficaz de estimar os valores genômicos de predição a partir de informações genéticas com base em métodos estatísticos, como a melhor predição linear imparcial (BLUP). O uso de haplótipos, agrupamentos de polimorfismo de nucleotídeo único (SNP) como marcadores, em vez de SNPs individuais, pode melhorar a precisão da predição genômica (Won et al. 2020). Sendo que, a probabilidade de um loco de característica quantitativa (QTL) estar em forte desequilíbrio de ligação (LD) com um agrupamento de marcadores é maior em comparação com um marcador individual (Won et al. 2020), obtendo maiores acurácias (Habier, Fernando e Dekkers, 2007; Wientjes, Veerkamp e Calus, 2013).

O LD refere-se a uma associação não aleatória entre os marcadores moleculares e os loci que controlam características quantitativas QTL (Goddard e Hayes, 2007; Yang *et al.*, 2010). O LD tem um efeito direto na genética das populações, por sua segregação não independente dos alelos dos diferentes locos, sendo influenciado por fatores naturais evolutivos ou aqueles que utilizamos no melhoramento genético. Estudos demonstram as variações do LD em algumas espécies de animais. Como o LD para aves varia de 0,32 a 0,73 ((Qanbari *et al.*, 2010), para suínos varia de 0,31 a 0,49 (Grossi *et al.*, 2017), para bovinos de corte, 0,45 (*bos taurus*), 0,25 (*bos indicus*), 0,32 (mestiços) (Porto-Neto, Kijas e Reverter, 2014) e para bovinos de leite da raça holandesa varia de 0,20 a 0,72(Mokry *et al.*, 2014).

Um haplótipo define uma região do genoma que compreende um conjunto de marcadores genéticos vizinhos, em que seus alelos em fases são provavelmente herdados conjuntamente (Hess et al. 2017). A utilização de haplótipos para a predição genômica tem como finalidade aproveitar o maior LD e reduzir o número de variáveis sem a perda de informações (Cuyabano et al. 2014). Ao decorrer das gerações, o LD vai diminuindo sendo este fato uma preocupação sobre a eficiência da seleção genômica. Cuyabano et al. (2014) propõem a abordagem de haplótipos para a predição genômica em alternativa ao uso de SNP. Isto porque eles supõem que os haplótipos estão em maior desequilíbrio de ligação (LD) com loci de características quantitativas do que marcadores únicos.

A princípio, foi proposta a ideia em que maiores densidade de marcadores garantiria que cada QTL controlando uma característica estaria em LD forte com pelo menos um marcador, levando a melhores acurácias (Meuwissen and Goddard 2010). Werner et al. (2018) concluíram que se obtêm melhores acurácias quando se utiliza conjunto de marcadores com menores densidades. Lorenz et al. (2010) afirmaram que podem existir vários motivos para

agrupar SNPs em blocos de haplótipos, e que as possíveis vantagens podem depender de fatores como arquitetura genética dos caracteres, padrões de LD na população de estudo e densidade de marcadores. VanRaden et al. (2013) complementam mostrando que é necessário não só genotipagem mais densa como também o melhor uso da mesma, e Prakapenka et al. (2020a) mostram que os modelos de predição utilizando haplótipos abrem muitas possibilidades para melhorar a acurácia da seleção genômica, porém requer um maior processamento de dados e tempo de computação do que modelos de predição de SNP único.

Além do exposto, as metodologias usadas para a construção e análises dos haplótipos podem influenciar nas acurácias de predição. Assim, há poucos estudos baseados na melhor metodologia para a construção e análises dos haplótipos, a quantidade de SNPs a serem usados para formar um haplótipos, bem como a estrutura genética da população. Em seu estudo, Jónás et al. (2017) avaliaram populações de bovinos da raça Montbéliarde e propuseram um pipeline simples que incorpora simultaneamente desequilíbrio de ligação e informações de frequência alélica na avaliação genômica. Este trabalho teve como conclusão que o método proposto era, de fato, vantajoso e que a acurácia da avaliação genômica poderia ser melhorada.

Villumsen et al. (2009), em dados simulados para um ambiente típico de gado leiteiro, encontraram maior confiabilidade para haplótipos com 10 SNPs utilizando caracteres com herdabilidades iguais a 0,02 e 0,30. Calus et al. (2008) avaliaram, em dados simulados para animais bovinos, modelos de efeito de cada alelo de SNP e haplótipos construído a partir de 2 ou 10 marcadores, incluindo a covariância entre os haplótipos, combinando desequilíbrio de ligação e análise de ligação para um caráter com herdabilidade de 0,10 e 0,50, obtiveram pequenas diferenças entre os modelos para a característica de herdabilidade igual a 0,10, enquanto para o caráter com herdabilidade igual a 0,50, o modelo de haplótipos construídos a partir de 10 marcadores rendeu maiores acurácias dos valores.

O comprimento de blocos de haplótipos constitui um problema a ser estudado em diferentes níveis de desequilíbrio de ligação. Neste contexto, blocos mal construídos podem gerar informações incompletas sobre a variabilidade genética dentro da região analisada (Lin, Chakravarti e Cutler, 2004). Com objetivo de contornar esse problema, Li et al. (2007) propuseram uma análise de associação de haplótipos fundamentada em uma estrutura de janela deslizante de tamanho variável, a qual emprega análise de regressão regularizada resolvendo o problema de vários graus de liberdade do teste de haplótipos. Esta análise não exige um conhecimento prévio da estrutura do desequilíbrio de ligação para formação dos blocos.

Em resumo, nosso estudo concentrou-se em comparar a eficiência de SNPs únicos e de blocos de haplótipos sob o contexto de predições genômicas em gerações avançadas no melhoramento genético, em duas populações com diferentes níveis de LD.

2. OBJETIVOS

O objetivo desse trabalho é avaliar a eficiência da seleção genômica em dois softwares com uma mesma população, avaliar a eficiência da seleção genômica baseada em duas populações com diferentes níveis de LD. Comparando a eficiência de marcadores polimorfismo de nucleotídeo único e blocos de haplótipos com diferentes tamanhos de blocos em um contexto de gerações avançadas de predição.

3. MATERIAIS E MÉTODOS

3.1. CONJUNTO DE DADOS

Foi utilizado o software REALbreeding (disponível mediante solicitação) para empregar a simulação de duas populações, sendo a primeira população com menor LD e a segunda população com um LD maior. Para cada população, foi realizada uma simulação dos genótipos e fenótipos de 5000 genitores fundadores (geração 1) e das próximas cinco gerações, cada qual com 5.000 indivíduos, totalizando 30.000 indivíduos. Os pais de cada geração subsequente vieram da geração anterior, sem seleção, originando progênies de irmãos completos de tamanhos diferentes, mas mantendo o tamanho da próxima geração constante em 5.000 indivíduos.

Com base em nossas informações, o REALbreeding distribuiu aleatoriamente 60.000 SNPs, 400 genes menores em 10 cromossomos (6.000 SNPs e 40 genes por cromossomo). A densidade média foi de um SNP a cada 0,01 cM. A característica simulada terá valores genotípicos mínimo e máximo para homozigotos de 0 e 2, respectivamente. Esses valores foram usados para calcular os desvios a (a diferença entre o valor genotípico do homozigoto de maior expressão e a média dos homozigotos) e d (desvio de dominância). O desvio de dominância foi calculado apartir do grau de dominância (d / a).

Definimos dominância positiva ($0 < d / a \leq 1{,}2$). Os verdadeiros valores genéticos aditivos e de dominância foram calculados a partir das frequências de genes da população, valores de LD, efeitos médios de substituição de genes e desvios de dominância. Os valores fenotípicos foram calculados a partir da média real da população, valores aditivos e de dominância e de efeitos de erro amostrados de uma distribuição normal. A variância do erro foi calculada a partir da herdabilidade no sentido amplo, assumida como sendo 0,30.

A construção de haplótipos é possível a partir de diferentes metodologias. Cada uma delas apresenta vantagens e desvantagens de acordo com a população estudada e com a característica e distribuição dos marcadores ao longo do genoma. A escassez de estudos com finalidade de definir a metodologia mais adequada para construção a avaliação de haplótipos resulta em uma falta de consenso na literatura em relação à quantidade de SNPs que devem formar um haplótipo, além de qual método se apresenta com melhor desempenho para construção dos conjuntos de SNPs (Paulista *et al.*, 2018). Assim, foi realizada uma análise de LD em ambos os cenários, com diferentes metodologias de construção de haplótipos, como *Gabriel et. Al, four gamete rule e solid spline*, a fim de compreender melhor os resultados e apoiar as comparações entre SNP ou modelo baseado em haplótipos. Assim, o cromossomo 1 da segunda geração usada para fazer as predições, foi escolhido para realizar a análise LD utilizando o software Haploview (Barrett et al. 2005b). Para a população com menor LD, obtivemos padrão de blocos de em média 2 a 3 SNPs (apresentado na figura 1) e para a população com maior LD os estudos sugeriram utilizar blocos variando de 2 a 4 SNPs (apresentado na figura 2), porém decidimos seguir o padrão de 4 para acompanhar os estudos e o 10 SNPs por blocos para acompanhar a literatura.

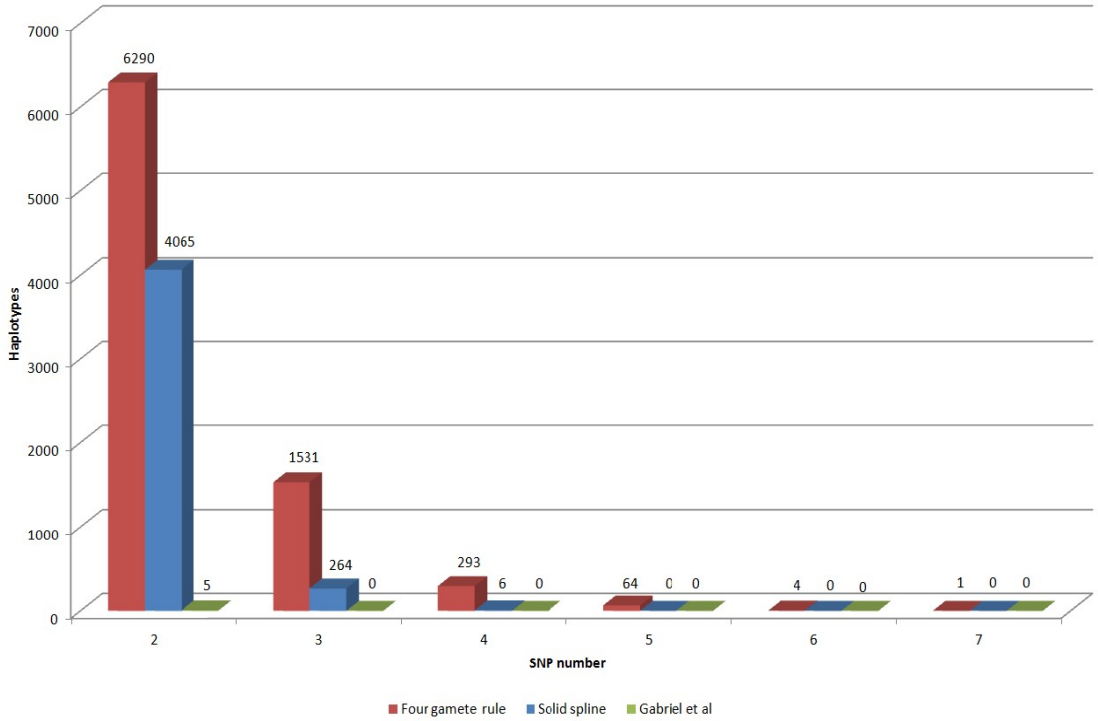

Figura 1 Distribuição dos blocos de haplótipos com diferentes metodologias de construção na população com menor LD

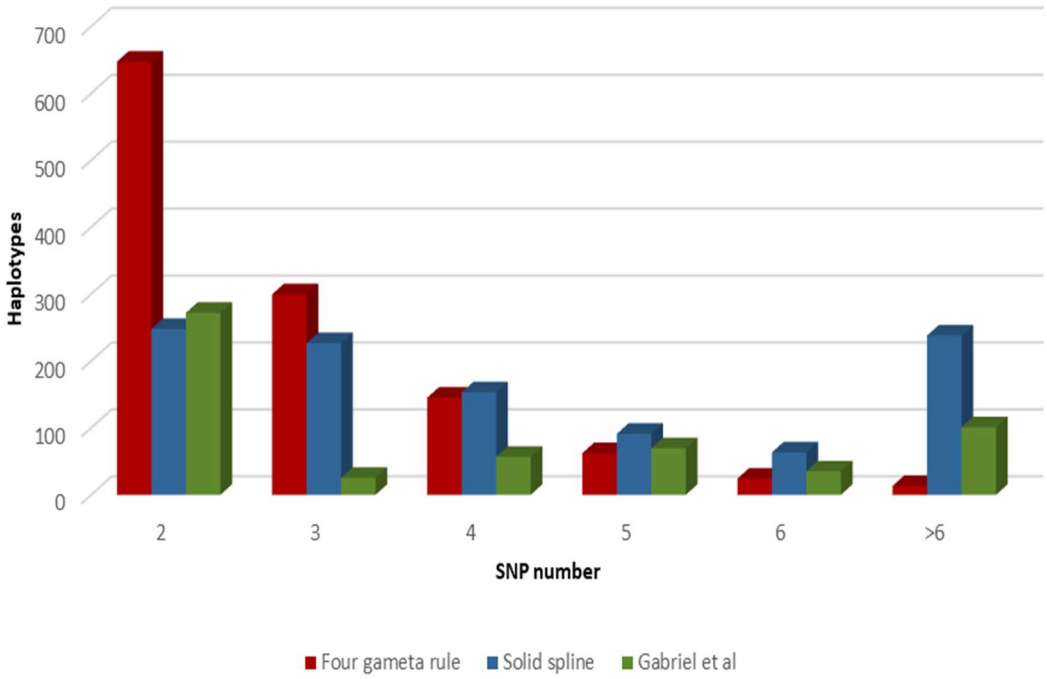

Figura 2 Distribuição dos blocos de haplótipos com diferentes metodologias de construção na população com maior LD.

3.2. ANALISES NO SOFTWARE R EM POPULAÇÃO DE MENOR LD

A chamada de haplótipos foi feita por cada cromossomo agrupando 10 SNPs por vez para formar o bloco de haplótipos. Assim, os diferentes alelos de haplótipos dentro de cada bloco foram pontuados para formar os genótipos de alelos de haplótipos de amostra com base no número de cópias (ou seja, 0, 1 ou 2 cópias) para cada indivíduo diplóide (apenas dois alelos para cada um). Em seguida, os genótipos dos alelos dos haplótipos de cada cromossomo foram reunidos após a aplicação dos critérios de frequência do alelo alternativo (MAF) maior que 0,01 para realização da análise. O pacote GHap R (Utsunomiya et al. 2017) foi usado para obter os alelos do haplótipo de dados SNP em fases. Para a obtenção dos dados da fase a partir dos SNP, foi utilizado o software AlphaImpute versão 1.9.8 (Hickey et al. 2011), fornecendo ao software as informações genealógicas de todas as gerações.

As análises estatísticas foram feitas com o pacote sommer do R (Covarrubias-Pazaran 2016). Apenas SNPs com MAF maior que 0,05 foram usados na análise. As acurácias da predição de valor aditivo foram calculadas para todos os indivíduos a serem preditos. Também, a coincidência dos 500 melhores indivíduos (10% do tamanho da população) foi calculada após classificar os indivíduos preditos e seu valor genético aditivo correspondente a partir dos dados simulados. Assim, os primeiros 500 indivíduos em cada um desses dados classificados foram comparados. A coincidência paramétrica, usada para fazer comparações,

foi calculada a partir dos valores fenotípicos e genéticos aditivos classificados dos dados simulados.

O tamanho do conjunto de treinamento foi de 2.000 indivíduos amostrados aleatoriamente da geração fundadora (geração 1). Os pais fundadores não tiveram nenhuma relação genética entre eles. Este procedimento foi repetido 50 vezes para cada geração a ser prevista. A predição da acurácia do valor aditivo foi calculada como a correlação média entre os valores aditivos verdadeiros calculados por REALbreeding e os valores previstos por GBLUP na validação independente definida ao longo de 50 reamostragem. Como um limite para a acurácia máxima alcançável, realizamos a predição genômica dentro de cada geração empregando 5.000 indivíduos como população de validação.

3.2.1. MODELAGEM ESTATÍSTICA

A partir da matriz de genótipos de alelos de haplótipos para todos os indivíduos, a matriz de relacionamento genômico foi construída pela expressão:

$$\boldsymbol{G} = \boldsymbol{qMDM'}$$

em que $\boldsymbol{M}$ é a matriz $(n \times h)$ centrada do genótipo dos alelos do haplótipo, onde n é o número de indivíduos e h é o número dos alelos do haplótipo, $\boldsymbol{D} = \text{diag}(d_i)$, em que d_i é o peso do alelo do haplótipo i (padrão $d_i = 1$), e $\boldsymbol{q} = \text{tr}(\boldsymbol{MDM'}) - 1\,n$ (REFERÊNCIA). A matriz de relacionamento genômico derivada de um único SNP foi a matriz proposta por Endelman and Jannink (2012).

Ajustamos um modelo BLUP genômico (GBLUP), para predição baseada em SNP ou haplótipos, dado por:

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{Z\alpha} + \boldsymbol{\varepsilon}$$

em que $\boldsymbol{y}$ é o vetor de valores fenotípicos, μ é a média da população, $\boldsymbol{\alpha}$ é o valor genético aditivo do indivíduo, $\boldsymbol{Z}$ é a matriz de incidência e $\boldsymbol{\varepsilon}$ é o vetor de erro. A variância de $\boldsymbol{\alpha}$ é $\boldsymbol{G\sigma_a^2}$, em que $\boldsymbol{G}$ é a matriz de relacionamento genômico aditivo e $\boldsymbol{\sigma_a^2}$ é a variância aditiva.

3.3. ANÁLISES NO SOFTWARE GVCHAP

O programa GVCHAP (Prakapenka et al. 2020) foi usado para a análise das duas populações. Para os blocos de 4 SNP s por bloco. A chamada de haplótipos foi feita por cada cromossomo agrupando 4 SNPs por vez para formar o bloco de haplótipos. Assim, os diferentes alelos de haplótipos dentro de cada bloco foram pontuados para formar os genótipos de alelos de haplótipos de amostra com base no número de cópias (ou seja, 0, 1 ou

2 cópias) para cada indivíduo diploide (apenas dois alelos para cada um). Este programa implementa como matrizes de relacionamento genômico utilizando o método de Hayes-Goddard (Wang et al. 2014) que representa uma matriz genômica utilizando a média das diagonais dos elementos como denominador, permitindo a produção de estimativas de variância e herdabilidade na população de estudo que podem ser uma população aleatória ou consanguínea (Da 2015; Prakapenka et al. 2020).

Foram realizadas 10 simulações para calcular a acurácia da predição com observações dos valores aditivos. Para cada simulação, 2.000 indivíduos foram selecionados aleatoriamente entre os 5.000 indivíduos da geração fundadora, cujos indivíduos não apresentavam nenhuma relação de parentesco entre eles, como representação da população treinamento. Para toda a população de validação, os valores aditivos foram omitidos no cálculo do GBLUP (Liang et al. 2020). O programa GVCHAP calculou a melhor predição linear imparcial genômica de valores genéticos (GBLUP) de efeitos aditivos e o valor genotípico total é calculado usando as estimativas da última das iterações de máxima verossimilhança genômica restrita (GREML) (Prakapenka et al. 2020). O mesmo foi feito para os blocos de 10 SNPs, alterando apenas o número de SNPs por blocos. E também foi realizado para ambas as populações.

4. RESULTADOS E DISCUSSÃO

Como previsto, não houve grandes diferenças nas análises realizadas pelos softwares. Feitas as análises com a população de menor LD, ambos os softwares resultaram em uma maior acurácia para análises com SNP individual quando comparados com blocos de haplótipos com 10 SNPs (Tabela 1 e Tabela 3).

De modo geral, uma população de treinamento com maior grau de parentesco com a população de validação resulta em uma maior acurácia na predição de valores, e o maior grau de LD nas populações favorece as análises por haplótipo. O grau de LD vai reduzindo após gerações. Assim, as análises via SNP como marca únicas passam a ser favorecidas em ambas as populações. (Tabela 1 a 5).

Para as análises feitas no software R, para a população com menor nível de LD e blocos de haplótipos igual a 10 SNP, a acurácia da predição e a coincidência dos melhores indivíduos da população com base em marcadores SNP únicos são superiores em relação ao uso de haplótipos. Obtendo para o modelo baseado em SNP individual acurácias de predição variando de 0,3891 a 0,4864 para a geração 6 e 2, respectivamente. Para o modelo baseado em haplótipo, essas acurácias variaram de 0,2515 a 0,3886 para a geração 6 e 2,

respectivamente. Em geral, as acurácias com base em haplótipos (0,3089) são 30% mais baixas do que com base em SNPs (0,4404). A coincidência dos melhores indivíduos (10%) variou de 0,2466 a 0,3859 para o modelo baseado em SNP e de 0,1959 a 0,3349 para o modelo baseado em haplótipo para a geração 6 e 2, respectivamente. Por coincidência, o modelo baseado em haplótipos forneceu resultados 21% mais baixos em média. Comparando a média de coincidência paramétrica de 0,3156, o modelo baseado em SNP forneceu 7,07% menos coincidência e o modelo baseado em haplótipo 26,93% menos. Há uma tendência de redução, tanto na acurácia das predições quanto na coincidência, ao longo das gerações. Portanto, em nossa investigação, com o tamanho do populações de treinamento, modelo, herdabilidade e outros fatores mantidos constantes, a seleção genômica não supera o melhor resultado alcançável com a seleção fenotípica (Tabela 1).

Foram feitas análises no software GVCHAP, para a população de menor LD com SNP único, blocos de haplótipos igual a 4 e 10 SNPs, obtivemos os seguintes resultados. Nas análises feitas com blocos de haplótipos igual a 4 SNPs, o modelo baseado em SNP único forneceu acurácias de predição variando de 0,57 a 0,40 e o modelo baseado blocos de haplótipos com 4 SNPs forneceu a acurácia de predição variando de 0,49 a 0,13 (tabela 2) e nas análises realizadas com blocos de haplótipos igual a 10 SNPs, os valores da acurácia para os blocos de haplótipos variaram de 0,44 a 0,29 e o modelo baseado e SNP individual variou de 0,57 a 0,52 (tabela 3).

Para a população com maior nível de LD, a análise da acurácia da predição utilizando SNP único e blocos de haplótipos de 4 e 10 SNPs, está apresentado na Tabela 4 e 5. No cenário comparando SNP único e blocos de haplótipos com 4 SNPs, obtivemos as acurácias de predição com base no SNP variam de 0,1435 a 0,3971 para a população 5 e 1 respectivamente, e com base em haplótipos os valores variam de 0,1380 a 0,5648 para a população de 5 a 1 respectivamente. Em geral, as acurácias com base em SNP são 40% mais baixas do que haplótipos. Adquirimos resultados similares nas análises da acurácia utilizando SNP único e blocos de haplótipos com 10 SNPs, as acurácias de predição com base no SNP variaram de 0,01354 a 0,4036 para população 6 e 1 respectivamente, e com base em haplótipos os valores variaram de 0,2114 a 0,5852 para a população de 6 e 1 respectivamente. Em geral, as acurácias com base em SNPs são 23% mais baixas do que com base em haplótipos. Na geração fundadora, a acurácia para SNP foi em média de 0,4036 e para haplótipos 0,5852. Na geração 2, para os SNPs foram 0,3810 e para haplótipos 0,567626. Os valores decaíram no ponto em que, na geração 6, o valor de acurácia usando SNPs (0,1354) superou a acurácia para haplótipos (0,2114). Da geração 2 à geração 4, o declínio médio na

acurácia da predição de indivíduos foi de 20% usando marcadores SNP únicos no modelo, e de 45% usando haplótipos. Da geração 5 à geração 6, houve um declínio médio de 85% para marcadores SNP únicos no modelo, e de 80% para haplótipos. Nestas últimas, os valores da acurácia por SNPs superaram os valores das acurácias para haplótipos.

Tabela 1 Análises utilizando o Software R para acurácia de predição e coincidência (desvio padrão) da população de baixo LD em função do modelo usado, geração de validação e tamanho de validação, usando blocos de haplótipo com 10 SNPs.

Densidade do marcador	Modelo	Geração	Amostra	Acurácia	Coincidência *ref=0.3156
60,000	SNP	2	3000[a]	0.4864 (0.0185)	0.3859 (0.0243)
		3	5000	0.4765 (0.0136)	0.3086 (0.0191)
		4	5000	0.4357 (0.0155)	0.2649 (0.0180)
		5	5000	0.4143 (0.0148)	0.2606 (0.0159)
		6	5000	0.3891 (0.0158)	0.2466 (0.0180)
	Haplótipo	2	3000[a]	0.3886 (0.0180)	0.3349 (0.0214)
		3	5000	0.3480 (0.0200)	0.2364 (0.0201)
		4	5000	0.3038 (0.0251)	0.2004 (0.0175)
		5	5000	0.2528 (0.0276)	0.1853 (0.0175)
		6	5000	0.2515 (0.0192)	0.1959 (0.0171)

*: Valores paramétricos

[a]: Validação independente com a 2 geração

Tabela 2 Análise utilizando o software GVCHAP para acurácia de predição da população de menor LD em função do modelo usado, geração de validação e tamanho de validação, usando blocos de haplótipo com 4 SNPs.

Densidade do marcador	Modelo	Geração	Amostra	Acurácia
60,000	SNP	1	3000	0.569172(0.004)
		2	5000	0.568272(0.007)
		3	5000	0.564288(0.005)
		4	5000	0.482829(0.032)
		5	5000	0.470567(0.005)
		6	5000	0.404739(0.258)
	Haplótipo	1	3000[a]	0.490129(0.011)
		2	5000	0.490381(0.010)
		3	5000	0.479067(0.005)
		4	5000	0.525066(0.417)
		5	5000	0.269201(0.003)
		6	5000	0.131149(0.185)

Tabela 3 Análise utilizando o software GVCHAP para acurácia de predição da população de menor LD em função do modelo usado, geração de validação e tamanho de validação, usando blocos de haplótipos com 10 SNPs.

Densidade do marcador	Modelo	Geração	Amostra	Acurácia
60,000	SNP	1	3000	0.567486 (0.001)
		2	5000	0.567109 (0.006)
		3	5000	0.564840(0.005)
		4	5000	0.550982(0.019)
		5	5000	0.530980(0.017)
		6	5000	0.524010(0.125)
	Haplótipo	1	3000	0.461802(0.028)
		2	5000	0.453464(0.016)
		3	5000	0.460122(0.005)
		4	5000	0.254798(0.015)
		5	5000	0.211452(0.035)
		6	5000	0.1284345(0.110)

Tabela 4 Análise utilizando o software GVCHAP para acurácia de predição da população de maior LD em função do modelo usado, geração de validação e tamanho de validação, usando blocos de haplótipos com 4 SNPs.

Densidade do marcador	Modelo	Geração	Amostra	Acurácia
60,000	SNP	1	3000	0.397183 (0.003)
		2	5000	0.384792 (0.019)
		3	5000	0.396342 (0.005)
		4	5000	0.254024 (0.082)
		5	5000	0.143564 (0.103)
		6	5000	0.147359 (0.193)
	Haplótipo	1	3000	0.564818 (0.014)
		2	5000	0.552442 (0.028)
		3	5000	0.571382 (0.008)
		4	5000	0.263979 (0.088)
		5	5000	0.138052 (0.130)
		6	5000	0.141600 (0.186)

Tabela 5 Análise utilizando o software GVCHAP para acurácia de predição da população de maior LD em função do modelo usado, geração de validação e tamanho de validação, usando blocos de haplótipo com 10 SNPs.

Densidade do marcador	Modelo	Geração	Amostra	Acurácia
60,000	SNP	1	3000	0.403651 (0.002)
		2	5000	0.381013 (0.008)
		3	5000	0.398340 (0.008)
		4	5000	0.241745 (0.027)
		5	5000	0.215097(0.024)
		6	5000	0.135421 (0.177)
	Haplótipo	1	3000	0.585252(0.011)
		2	5000	0.567655 (0.006)
		3	5000	0.570422 (0.007)
		4	5000	0.245685 (0.007)
		5	5000	0.21145 (0.010)
		6	5000	0.128434 (0.169)

Observamos que explorar o LD entre marcadores, como no caso dos haplótipos, para realizar predição genômica não garante melhores resultados de seleção do que aqueles obtidos por SNPs individuais. Autores com Frischknecht et al. (2016) trabalharam com diferentes formas de construção de blocos de haplótipos, com base em comprimento físico em pares de bases (bp) ou em medidas de LD e concluíram que houve pequena diferenças entre os cenários estudados. Cuyabano et al. (2014) não encontraram nenhuma diferença entre o uso de haplótipos ou SNPs únicos quando a característica investigada tinha uma herdabilidade muito baixa (0,04). Em contrapartida, eles encontraram uma diferença significativa na acurácia da predição quando usou o LD médio limite superior a 0,45 entre quaisquer dois marcadores para definir os blocos de haplótipos. Jan et al. (2019) afirmaram que estudo de haplótipo tem o poder de melhorar a eficiência das análises no melhoramento genético quando se trabalha com LD superior a 0,50. Sallam et al. (2020) afirmaram que haplótipos podem melhorar a predição da acurácia genômica sobre SNPs únicos porque os haplótipos podem capturar melhor o desequilíbrio de ligação e a similaridade genômica, trabalharam com linhagens avançadas e cultivares da cultura trigo. Assim, observamos que vários autores também relatam que a acurácia na seleção genômica depende de outros fatores além do modelo de analises, seja ela por haplótipos ou SNP individual

Em nosso estudo, trabalhamos com duas populações com diferentes níveis de LD, nas quais observamos que o haplótipo garante melhores respostas nas acurácias das predições

quando a população possui um alto LD, como observados nos trabalhos citamos acima. Porém, observamos que o haplótipo não é capaz de explorar melhor o LD na população quando comparados a modelos de marcas únicas ao longo das gerações (com a queda do LD) nem quando trabalhamos com população com baixo LD. Autores confirmaram que mesmo haplótipos tendo vantagens frente ao modelo de marcas únicas, eles requerem um maior processamento de dados e tempo de computação, além de sua eficiência está altamente ligada com a estrutura genética da população (Prakapenka et al. 2020).

Diferentes formas de obter os haplótipos resultam em diferentes eficiências de uso na predição genômica. A matriz de relacionamento genômico tradicional considera LD completo entre SNPs e QTL, mas ignora LD entre SNPs, especialmente em regiões menores (Habier, Fernando e Garrick, 2013). Jónás et al. (2016) investigaram métodos para construir haplótipos após uma análise prévia dos efeitos dos marcadores individualmente. Os autores obtiveram os haplótipos mesclando SNPs flanqueados em torno daqueles de maior efeito sobre a característica em um determinado tamanho de haplótipo (3, 4, 5 SNPs???) ou aplicando um limite para a frequência do alelo. Uma predição de aumento médio de 2% com base em haplótipos em vez de SNP único foi encontrada por eles. Ma et al. (2016) estudaram o caráter rendimento de grão em plantas de soja, e a acurácia da predição com base em marcadores selecionados com uma abordagem baseada em análises de blocos de haplótipos aumentou aproximadamente 4% em comparação com a amostragem de marcadores aleatória ou equidistante. Concluíram, então, que a aplicação de pré-seleção de marcadores com base em blocos de haplótipos é uma opção interessante para uma implementação econômica da seleção genômica para rendimento de grãos no melhoramento de soja.

Mathew et al. (2018) encontraram apenas uma melhora de 0,75% nas acurácias de predição usando uma matriz de relacionamento corrigida para LD quando comparada à matriz genômica SNP tradicional com conjunto de dados reais de populações de milho, arroz, camundongos e gado. Por outro lado, Uemoto et al. (2017) encontraram acurácias de predição ligeiramente maiores para o modelo BLUP empregando matriz de relacionamento genômico construída a partir de SNPs em comparação à matriz construída a partir de haplótipos, da mesma forma nossos resultados. Ferdosi et al. (2016), empregando três métodos diferentes para construir a matriz de relacionamento genético entre indivíduos de bovinos de corte, também encontraram melhores acurácias de predição para haplótipo baseado em modelo baseado em SNP único. Esta superioridade foi de cerca de 6,51% usando um conjunto de dados reais de linhagens de porcos Duroc. Os autores afirmaram sobre a propriedade da matriz de relacionamento baseada na captura de relacionamentos muito antigos do SNP,

enquanto a matriz de relacionamento baseada em haplótipos captura relacionamentos de idade intermediária.

Vários fatores influenciam nos resultados de estudos de predição genômica, dentre eles o comprimento dos blocos. Neste trabalho utilizamos para a população de baixo LD o comprimento de 10 marcadores SNP e os blocos de haplótipos sem sobreposição, sendo considerados adequados para construir os alelos de haplótipos com base em estudos anteriores (Villumsen et al. (2009); Ferdosi et al. (2016)). Esses autores demonstraram que, se a semelhança de haplótipos é a base para determinar relações, aumentar o número de alelos de haplótipos, como acontece com a adoção de blocos de haplótipos maiores, geralmente resultará em relacionamentos mais baixos dentro de qualquer segmento. Além disso, Ferdosi et al. (2016) concluíram que haplótipos mais longos resultaram em uma matriz de relacionamento semelhante a uma matriz de identidade. Em contrapartida, para definição do tamanho de blocos para a população de alto LD em nossa investigação, utilizamos blocos de 4 SNPs, número que está de acordo com os estudos de blocos realizados pelo software Haploview (Barrett *et al.*, 2005), além de estudos anteriores afirmarem este comprimento como adequado nas análises. HAYES et al. (2007) compararam a acurácia de seleção assistida por marcador (MAS) usando marcadores únicos ou haplótipos de marcadores em um conjunto de dados de bovinos Angus. Concluíram que a precisão do MAS aumentou conforme o número de marcadores no haplótipo em torno do QTL aumentou, em blocos de haplótipos com 4 ou mais SNPs e afirmaram que a acurácia aumentou quando houve um LD mais alto entre os marcadores com o QTL.

Assim, com base nos estudos anteriores, a construção de alelo de haplótipo apropriado faz-se necessária para capturar a relação genética entre os indivíduos para prever valores genéticos para realizar a seleção adicional. No entanto, a verdadeira utilidade dos haplótipos na seleção genômica ainda não é bem compreendida. Dessa forma, nosso trabalho fez o uso do software Haploview (Barrett *et al.*, 2005) como base para análises do perfil de desequilíbrio da população e como fonte para definir o comprimento do bloco de haplótipos. Jónás et al. (2016) descobriram que os haplótipos contendo aqueles SNPs previamente analisados (efeitos maiores na expressão fenotípica) fornecem um aumento na acurácia de predição do que o modelo SNP único, variando entre 0,8 e 2,9% para cinco características de um conjunto de dados de gado leiteiro real. Os estudos diferem entre si quanto à melhor estratégia a ser empregada considerando o uso de haplótipos. Calus et al. (2009), por exemplo, não encontraram diferenças significativas na eficiência dos haplótipos para predição variando o comprimento do haplótipo de 2 a 20 SNPs. Frischknecht et al. (2016) investigaram

diferentes definições de haplótipos para prever quatro características no conjunto de dados de gado marrom suíço variando o comprimento físico ou medida de LD para construir os haplótipos. Os autores afirmaram que apenas melhorias marginais foram encontradas comparando as acurácias obtidas dos haplótipos versus 50K SNP. Também empregando matriz de relacionamento derivada de um único SNP ou haplótipos, como feito em nosso estudo, Karimi et al. (2018), avaliaram a acurácia de predição para 57 caracteres de Holstein. Esses autores também construíram o bloco de haplótipos considerando um número fixo de SNPs adjacentes (5, 10, 15 ou 20) e encontraram diferenças nas acurácias do modelo baseado em SNP e no modelo baseado em haplótipo variando de -4,2% a +3,3%. Para a maioria das características com baixa herdabilidade, o modelo baseado em haplótipos teve desempenho ruim do que o baseado em SNP. O comprimento do haplótipo de 20 SNPs resultou no maior viés de predição e nas menores precisões, enquanto o comprimento de 5 e 10 SNPs foram semelhantes.

Villumsen et al. (2009), apontaram as perspectivas da seleção genômica em um contexto de gado leiteiro utilizando haplótipos, mantendo altos valores de confiabilidade até sete gerações. Esses autores tiveram maior confiabilidades para blocos de haplótipos com 10 SNPs. Esses autores simularam sete gerações de uma população de gado leiteiro, sem seleção, como feito aqui, e encontraram uma superioridade geral das predições baseadas em haplótipos sobre os SNPs individuais trabalhando com característica com uma herdabilidade variando de 0,02 a 0,30. Em nossas investigações, nas primeiras gerações da população de alto LD, a acurácia das predições via haplótipos superam as análises por SNPs. Porém, ao longo das gerações com o decréscimo do LD, as análises por haplótipo perdem o seu poder até chegar ao ponto em que as análises por SNPs superam com diferenças relativamente muito baixas as análises por haplótipos, como foi observado na população de baixo LD. Muitos estudos não confirmam uma vantagem clara do uso de haplótipos para predição, da mesma forma que nossos resultados apontaram. Contudo, nossos estudos não apontam para o uso de haplótipos para predição ao longo de gerações em estruturas genéticas populacionais de baixo LD. No entanto, ainda é necessário fazer maiores estudos para o uso de haplótipos para predição em gerações futuras.

5. CONCLUSÕES

Assim, com base em nossa investigação podemos afirmar que: 1) O software utilizado não influencia na conclusão a respeito do modelo a ser utilizado. 2)A acurácia da predição do SNP de um caráter quantitativo deve ser maior do que a acurácia da predição do haplótipo se houver menor LD; 3) A acurácia da predição haplótipo em estrutura genética de maior LD é mais eficiente quando houver maior grau de parentesco; ao reduzir o grau de parentesco, a acurácia se reduz ao ponto que as análises por SNPs se tornam mais vantajosas.

REFERÊNCIAS

BARRETT, J. C. *et al.* Haploview: analysis and visualization of LD and haplotype maps. **Bioinformatics**, v. 21, n. 2, p. 263–265, 15 jan. 2005.

CALUS, M. P. *et al.* Effects of the number of markers per haplotype and clustering of haplotypes on the accuracy of QTL mapping and prediction of genomic breeding values. **Genetics Selection Evolution**, v. 41, n. 1, p. 11, 15 dez. 2009.

CALUS, M. P. L. *et al.* Accuracy of Genomic Selection Using Different Methods to Define Haplotypes. **Genetics**, v. 178, n. 1, p. 553–561, 1 jan. 2008.

ENDELMAN, J. B.; JANNINK, J.-L. Shrinkage Estimation of the Realized Relationship Matrix. **G3: Genes|Genomes|Genetics**, v. 2, n. 11, p. 1405–1413, nov. 2012.

FERDOSI, M. H.; HENSHALL, J.; TIER, B. Study of the optimum haplotype length to build genomic relationship matrices. **Genetics Selection Evolution**, v. 48, n. 1, p. 75, 29 dez. 2016.

FRISCHKNECHT, M. *et al.* Genomic Prediction using Haplotypes in Brown Swiss. **Interbull Bulletin**, v. 50, n. 50, p. 34–38, 2016.

GODDARD, M. E.; HAYES, B. J. Genomic selection. **Journal of Animal Breeding and Genetics**, v. 124, n. 6, p. 323–330, 7 dez. 2007.

GROSSI, D. A. *et al.* Genetic diversity, extent of linkage disequilibrium and persistence of gametic phase in Canadian pigs. **BMC Genetics**, v. 18, n. 1, p. 6, 21 dez. 2017.

HABIER, D.; FERNANDO, R. L.; DEKKERS, J. C. M. The Impact of Genetic Relationship Information on Genome-Assisted Breeding Values. **Genetics**, v. 177, n. 4, p. 2389–2397, dez. 2007.

HABIER, D.; FERNANDO, R. L.; GARRICK, D. J. Genomic BLUP Decoded: A Look into the Black Box of Genomic Prediction. **Genetics**, v. 194, n. 3, p. 597–607, 1 jul. 2013.

HAYES, B. J. *et al.* Accuracy of marker-assisted selection with single markers and marker haplotypes in cattle. **Genetical Research**, v. 89, n. 4, p. 215–220, 21 ago. 2007.

JAN, H. U. *et al.* Genome-wide haplotype analysis improves trait predictions in Brassica napus hybrids. **Plant Science**, v. 283, p. 157–164, jun. 2019.

JÓNÁS, D. *et al.* Alternative haplotype construction methods for genomic evaluation.

Journal of Dairy Science, v. 99, n. 6, p. 4537–4546, jun. 2016.
JÓNÁS, D.; DUCROCQ, V.; CROISEAU, P. Short communication: The combined use of linkage disequilibrium–based haploblocks and allele frequency–based haplotype selection methods enhances genomic evaluation accuracy in dairy cattle. **Journal of Dairy Science**, v. 100, n. 4, p. 2905–2908, abr. 2017.
KARIMI, Z. *et al.* Assessing haplotype-based models for genomic evaluation in Holstein cattle. **Canadian Journal of Animal Science**, v. 98, n. 4, p. 750–759, 1 dez. 2018.
LI, Y.; SUNG, W.-K.; LIU, J. J. Association Mapping via Regularized Regression Analysis of Single-Nucleotide–Polymorphism Haplotypes in Variable-Sized Sliding Windows. **The American Journal of Human Genetics**, v. 80, n. 4, p. 705–715, abr. 2007.
LIN, S.; CHAKRAVARTI, A.; CUTLER, D. J. Exhaustive allelic transmission disequilibrium tests as a new approach to genome-wide association studies. **Nature Genetics**, v. 36, n. 11, p. 1181–1188, 24 nov. 2004.
MA, Y. *et al.* Potential of marker selection to increase prediction accuracy of genomic selection in soybean (Glycine max L.). **Molecular Breeding**, v. 36, n. 8, p. 113, 28 ago. 2016.
MATHEW, B.; LÉON, J.; SILLANPÄÄ, M. J. A novel linkage-disequilibrium corrected genomic relationship matrix for SNP-heritability estimation and genomic prediction. **Heredity**, v. 120, n. 4, p. 356–368, 14 abr. 2018.
MOKRY, F. B. *et al.* Linkage disequilibrium and haplotype block structure in a composite beef cattle breed. **BMC Genomics**, v. 15, n. S7, p. S6, 27 out. 2014.
PAULISTA, U. E. *et al.* Ampla Com Dados Simulados. 2018.
PORTO-NETO, L. R.; KIJAS, J. W.; REVERTER, A. The extent of linkage disequilibrium in beef cattle breeds using high-density SNP genotypes. **Genetics Selection Evolution**, v. 46, n. 1, p. 22, 24 dez. 2014.
PRAKAPENKA, D. *et al.* GVCHAP: A Computing Pipeline for Genomic Prediction and Variance Component Estimation Using Haplotypes and SNP Markers. **Frontiers in Genetics**, v. 11, n. April, p. 1–10, 2020.
QANBARI, S. *et al.* Linkage disequilibrium reveals different demographic history in egg laying chickens. **BMC Genetics**, v. 11, n. 1, p. 103, 2010.
SALLAM, A. H. *et al.* Improving Prediction Accuracy Using Multi-allelic Haplotype Prediction and Training Population Optimization in Wheat. **G3:**

Genes|Genomes|Genetics, v. 10, n. 7, p. 2265–2273, jul. 2020.

UEMOTO, Y. *et al.* Genomic evaluation using SNP- and haplotype-based genomic relationship matrices in a closed line of Duroc pigs. **Animal Science Journal**, v. 88, n. 10, p. 1465–1474, out. 2017.

VILLUMSEN, T. M.; JANSS, L.; LUND, M. S. The importance of haplotype length and heritability using genomic selection in dairy cattle. **Journal of Animal Breeding and Genetics**, v. 126, n. 1, p. 3–13, fev. 2009.

WIENTJES, Y. C. J.; VEERKAMP, R. F.; CALUS, M. P. L. The Effect of Linkage Disequilibrium and Family Relationships on the Reliability of Genomic Prediction. **Genetics**, v. 193, n. 2, p. 621–631, 1 fev. 2013.

YANG, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. **Nature Genetics**, v. 42, n. 7, p. 565–569, 20 jul. 2010.